International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2020.122415

Scopus

Research Article

IJCRR
Section: Healthcare
Sci. Journal Impact Factor: 6.1 (2018)
ICV: 90.90 (2018)

Diabetes Diagnosis in Population by Intelligible Machine Learning

Vidushi, Akash Rajak, Ajay Kumar Shrivastava

KIET Group of Institutions, Delhi-NCR, Ghaziabad-201206, India

ABSTRACT

Introduction: At this junction, machine learning demand is enhancing in almost every critical area to catch interesting and decision-making patterns. This inductive research objective is investigating sophisticated different techniques of machine learning to effectively analyze health data. Naturally, the present health-related dataset is most sensitive, crucial, and needs accurate analysis, hence result generated by different learning algorithms have paramount importance. This sensitivity enhances and promoted data analytics, interest, and role through machine learning in the health sector.

Objectives: This research aims to analyze and predict diabetes by applying elegant learning algorithms on the diabetes dataset. The article also shows a comparative study analysis of algorithms.

Methods: This research uses the median method to preprocess the dataset. After preprocessing, ten different machine learning algorithms are applied to the diabetes dataset in this paper.

Results: This document uses a diabetes dataset that has eight different symptoms or features to predict disease. To get a better classification technique, various ML mechanisms results are compared and analyzed. This study outcome can be further utilized in incoming research based on diabetic health problems.

Conclusion: A linear support vector machine shows better detection results compared to others.

Key Words: Machine Learning, Predictive Analysis, Gaussian Process, Diabetes Prediction, SVM, Decision Tree, Nearest Neighbor

INTRODUCTION

In the health sector, diabetes incidences are increasing globally at a fast pace and become a supreme concern. To detect or predict diabetes is a paramount health-related concern. Currently, to discover the required pattern ubiquitously Machine Learning (ML) is widely used. Health sector sensitivity accelerates researchers to work in this sector using machine learning. To work with data-driven problems machine learning is adopted in many areas. However, machine learning implementation for problems essentially needs data knowledge. There are several algorithms present in machine learning. These techniques show the different result with a different dataset. A dataset may have null values, incorrect information, and incomplete data. To get good result preprocessing of the dataset is required. Preprocessing improves performance and hence prediction can be done smoothly. This research uses the median method to preprocess the dataset. After preprocessing, eight different machine learning algorithms are applied to the diabetes dataset in this paper. This document also shows comparative study and analysis work performed on the dataset using 10 learning classifiers. Nearest Neighbor[1], Linear SVM[2], RBF SVM[3], Gaussian Process, Decision Tree, Random Forest[4], MLP Classifier, Adaboost, Naïve Bayes, and QDA machine learning classifiers are used in this research. These classifiers are implemented on the diabetes dataset and then their results are compared and analyzed. In this letter, these classifiers predict diabetes based on eight different features. Features included in this study are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. In medical science, diabetes is one of the chronic and uprising diseases that need special attention. Keeping the necessity of detecting diabetes problems in mind, researchers are giving their efforts towards this field.[5] Seriousness towards the health

Corresponding Author:

Vidushi, KIET Group of Institutions, Delhi-NCR, Ghaziabad-201206, India.
E-mail: vidushi.mca@kiet.edu or vidushi.mtech@gmail.com

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 06.07.2020 **Revised:** 19.09.2020 **Accepted:** 22.10.2020 **Published:** 14.12.2020

sector increases concern to predict or detect diabetes based on some features or symptoms. With the change of society in a way of living, eating, and physical work, the possibility of diabetes drastically rises. The Incurability of this disease makes its prediction more essential. However, proper diet, exercise, and precautions are needed to reduce and control complications and ill effects of the disease. The negative impact of this disease on the patient can lead to even death. If proper care and precautions are not taken on time then it can affect health badly.[6] This document works for a diabetes prediction based on eight different symptoms or features. This research predicts that either patient is diabetic or not based on the entire mentioned features. For prediction, this paper uses different machine learning algorithm and compare the potential of classifiers based on their accuracy. This article also elaborates and analyses comprehensively machine learning classification algorithms.

In this direction, work is going on continually. Daniel et al[7] computation techniques of parallel as well distributed system, also use techniques of deep learning for efficient and proper analysis of health care data. The paper focuses on the establishment of a relationship between variables of medical and laboratory assessment along with adverse event's occurrence. In health sector employing using deep learning.[8] Research provides relative merit analysis, technique pitfalls, along with future outlook. The main emphasis of this research is on key deep learning applications such as public health, medical imaging, etc.

Tian et al[9] analyzes data of trunk sway as well as techniques of machine learning to get automatically balance evaluation and provides assessment accurately outside the clinic. For mounting poisoning attacks a systematic approach as well as algorithm-independent across algorithms of machine learning and datasets in the health sector.[10] An approach for improving postprandial glucose regulation by using the ML-based KNN method is proposed.[11] The ML approach is useful in many applications. Authors Ref.[12] aim is investigating the purpose of sophisticated techniques of machine learning for personalized models' development that targets detecting risk in T2DM patients for non-fatal as well as fatal CVD incidence.

The aim of the authors Ref.[13] is assessing the association of type 2 diabetes, as well as HW phenotype and also predictive powers, are evaluated for combined Korean adult's TG levels with anthropometric measurements. PhysOnline is presented by researchers in paper.[14] PhysOnline is a pipeline that is built for Apache Spark that is an open-source platform to work for physiological data streaming to extract features online as well as through machine learning.

Wen et al[15] analyzed the detectability of Microaneurysms in this article with the use of pixel patches of size 25 by 25 extracted from the images of finds present in the database of DIAbeticRETinopathyi.e DIARETDB1.

MATERIALS AND METHODS

In the biomedical field, the most serious and critical disease even for human life is Diabetes[16]. Because health data is directly related to public life becomes critical and that's why the health industry needs additional and special attention. Health-related data are available vastly and it is growing continually and becomes much more complex. This enormous data open an opportunity for researchers to analyze it and based on that analyses try to predict or detect disease. This article uses the diabetes dataset and based on features available in this dataset predicts diabetes. Diabetes is one of the uprisings and most people affecting disease. It drastically affects life and plays a negative role in health or even can reduce an affected person's life. The incurable nature of this disease makes its prediction more vital. This disease can be handled by taking appropriate precautions and exercise timely and accurately. This is another important reason that also shows the necessity of this disease prediction.[16-18] This research applies prediction algorithms on the diabetes dataset and shows the result. This dataset is taken from Kaggle. This is a good dataset consist of eight different symptoms and 768 records. 768 different measurements are taken related to 8 different features, named pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age present in the dataset. Based on the mentioned symptoms disease is predicted. The main concern with any dataset is the presence of null values, non-relevant data, noisy data, etc. All these mentioned problems can lead to the problem of incorrect prediction which further leads the way towards the wrong judgment. And finally, the wrong decision goes to badly affect the complete system. To overcome this problem, preprocessing on a dataset is done.

Preprocessing is a way of getting the dataset with correct values that helps further in achieving better result compared to unprocessed data. Table. 1 briefly demonstrates pre-processing steps that can be used with any dataset and makes data efficient for further classification or prediction tasks. The dataset used in this research is the diabetes dataset. To predict diabetes leaning models trained by this dataset. Before applying learning algorithms, it is necessary to pre-process data.[19,20,21] After studying the dataset, it is observed that many cells have a value of zero. These zeros are nothing but null or missing values. This research replaces zero, error, and null values with median measurement. Table 2 and Figure 1 show all eight features of the dataset before pre-processing with several null values. After the implementation of the median, a dataset with null values is shown in Table 3.

The research uses a diabetes dataset of eight features and applies different machine learning models for analysis.[17,18] These models are Nearest Neighbor, Linear SVM, RBF SVM, Gaussian process, Decision tree, Random forest, MLP classifier, Adaboost, Naïve Bayes, and QDA. Table. 4 briefly

explains the learning models used in this research.[22-24]

RESULTS

This study is developed to detect diabetes early in a patient on the strength of eight different symptoms. So that required precautions at the right time can be taken. Diabetes is a disease which needs special attention because of the incurable nature it has. This crucial nature gives diabetes prediction paramount importance.

During the experiment, the result achieved to detect diabetes is shown in Table 5 and graphically analyze in the Figure 2. For the experiment, various learning algorithms are applied to the diabetes dataset. First of all, the diabetes dataset is preprocessed and then apply 10 different learning paradigms. The potential of these models is analyzed based on accuracy.

On account of the accuracy potential of learning, models are shown numerically in the Table 5 and using graphs in Figure 2. Using Table 5 and Figure 2 it is easy to understand and compare the results of each classifier. A confusion matrix of results is shown in Table 6.

DISCUSSION

The diabetes detection result is shown in the previous section. After a thorough analysis of Table. 4 and Figure. 2, it can be concluded that the linear support vector machine shows a better detection result compared to others. Because the accuracy of linear SVM[19]is higher than others proves that the potential of this classifier on the diabetes dataset is high[20]. Result concludes the following:

1. The linear support vector machine shows a good result for the small-size dataset and infeasible with a dataset having a large number of records[21]. This research dataset is not too large.
2. This classifier is good for a non-sparse dataset for binary classification[22]. The outcome label of the dataset used in this research is either diabetic or non-diabetic that is binary.
3. The classification output for this model is better for linear and this paper used a linear dataset.
4. This model can work well by identifying a small number of parameters[23]. This research dataset has eight features.
5. This technique is not working well in case of high computation[24] and the dataset of this research needs less computation.

CONCLUSION

A health-related issue like diabetes is one of the important diseases which needs paramount attention. The sensitivity of this disease enhances and promoted the interest in analyzing it through machine learning. This study analyses diabetes by using ML algorithms and found that the linear support vector machine shows better detection results compared to others.

ACKNOWLEDGEMENT

We acknowledge the immense help from scholars whose articles are cited and included in this manuscript. We are also grateful to authors/editors/publishers of all those articles, journals, and books from where literature for this article has been reviewed and discussed.

Conflict of Interest

We don't have any conflicts of interest.

Source of Funding

We don't have any funding.

REFERENCES

1. NirmalaDevi M, Balamurugan SA, Swathi UV. An amalgam KNN to predict diabetes mellitus. In2013 IEEE International Conference ON Emerging Trends in Computing, Communication, and Nanotech. (ICECCN) 2013 Mar 25;691-695.
2. Kaya GT. A hybrid model for classification of remote sensing images with linear SVM and support vector selection and adaptation. J Sel Top App Earth Obs Remote Sensing 2013 Apr;6(4):1988-97.
3. Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. J Biomed Health Inform 2014 May;19(2):728-34.
4. Tripoliti EE, Fotiadis DI, Manis G. Automated diagnosis of diseases based on classification: Dynamic determination of the number of trees in random forests algorithm. Transactions Info Tech Biomed 2011 Nov;16(4):615-22.
5. Kourouma KR, Apollinaire YA, ACKA FK. Health Insurance Coverage: A Cross-Sectional Study Among Patients Followedatthe Diabetes Centreof ABIDJAN COTE D'IVOIRE (CADA). Int J Curr Res Rev 2016 Mar;8(6):35.
6. Wagh SP, Bhagat SP, Bankar N, Jain K. Role of Vitamin-C Supplementation in Type II Diabetes Mellitus. Int J Curr Res 2020;12 (13), 61-64.
7. Sierra-Sosa D, Garcia-Zapirain B, Castillo C, Oleagordia I, Nuño-Solinis R, Urtaran-Laresgoiti M, et al. Scalable healthcare assessment for diabetic patients using deep learning on multiple GPUs. IEEE Trans Ind Inform 2019 May;15(10):5682-9.
8. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ. Deep learning for health informatics. J. Biomed Health Infor 2016;21(1):4-21.
9. Bao T, Klatt BN, Whitney SL, Sienko KH, Wiens J. Automatically evaluating balance: a machine learning approach. Trans Neural Syst Rehabil Eng 2019 Jan;27(2):179-86.
10. Mozaffari-Kermani M, Sur-Kolay S, Raghunathan A, Jha NK. Systematic poisoning attacks on and defenses for machine learning in healthcare. J Biomed Health Infor 2014 Jul 30;19(6):1893-905.

11. Aiello EM, Toffanin C, Messori M, Cobelli C, Magni L. Postprandial glucose regulation via KNN meal classification in type 1 diabetes. IEEE Control Syst Let 2018 Jun;3(2):230-5.
12. Zarkogianni K, Athanasiou M, Thanopoulou AC, Nikita KS. Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication. J Biomed Health Infor 2017;22(5):1637-47.
13. Lee BJ, Kim JY. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. J Biomed Health Infor 2015;20(1):39-46.
14. Sutton JR, Mahajan R, Akbilgic O, Kamaleswaran R. PhysOnline: an open-source machine learning pipeline for real-time analysis of streaming physiological waveform. J Biomed Health Infor 2018 May;23(1):59-65.
15. Cao W, Czarnek N, Shan J, Li L. Microaneurysm detection using principal component analysis and machine learning methods. Transac Nanobiosci 2018 May;17(3):191-8.
16. Rajesh R, Singh SA, Vaithy KA, Manimekalai K, Kotasthane D, Rajasekar SS. The effect of mucuna pruriens seed extract on pancreas and liver of diabetic Wistar rats. Int J Curr Res Rev 2016 Feb;8(4):61.
17. Rubaiat SY, Rahman MM, Hasan MK. Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. In 2018 International Conference on Inn. Engg. Techn (ICIET) 2018 Dec 27 (pp. 1-6).
18. Jaya R, Kumar SM. A Study on Data Mining Techniques Methods Tools and Applications in Various Industries. Int J Curr Res Rev 2016 Feb;8(4):34.
19. Barakat N, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. Transac Inform Tech Biomed 2010;14(4):1114-20.
20. Abbas H, Alic L, Rios M, Abdul-Ghani M, Qaraqe K. Predicting diabetes in healthy population through machine learning. In2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) 2019: 567-570).
21. Adeena KD, Remya R. Extraction of relevant dataset for support vector machine training: A comparison. In2015 International Conference on Advances in Computing, Communications, and Informatics (ICACCI) 2015 Aug 10: 222-227). IEEE.
22. Lazar A. A comparison of linear support vector machine algorithms on large non-sparse datasets. In2010 Ninth International Conference on Machine Learning and Appl. 2010 Dec 12 (pp. 879-882). IEEE.
23. Sebald DJ, Bucklew JA. Support vector machine techniques for nonlinear equalization. transactions on signal processing. 2000 Nov;48(11):3217-26.
24. Wang H, Hu D. Comparison of SVM and LS-SVM for regression. In2005 International Conference on Neural Networks and Brain 2005 Oct 13; 1: 279-283).

Table 1: Data Pre-Processing Steps

Step	Pre-Processing steps Names	Description
First Step	Data Cleaning	This step includes cleaning, missing, noisy, and other irrelevant values of data.
Second Step	Data Integration	Integrating data from heterogeneous sources to the same place.
Third Step	Data Transformation	It is a way to transforming or changing data into required format.
Forth Step	Data Reduction	In order to become cost efficient, reduce the data in small size.

Table 2: DatasetFeaturesand Number of Null Values (Before pre-processing)

Column/Feature Name	Number of Null values	% Null values
Pregnancies	111	14.453125
Glucose	5	0.651042
BloodPressure	35	4.557292
SkinThickness	227	29.557292
Insulin	374	48.697917
BMI	11	1.432292
Diabetes Pedigree Function	0	0
Age	0	0

Table 3: Dataset Features and Number of Null Values (After pre-processing)

Column/Feature Name	Number of Null Value
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
Diabetes Pedigree Function	0
Age	0

Table 4: Machine Learning Models Brief Description

Machine Learning Models	Explanation
Nearest Neighbor	Supervised learning method used to recognize patterns and works as per distance function
Linear Support Vector Machine	Data points are classified by finding the hyperplane in space
RBF (Radial basis function) Support Vector Machine	RBF is a Kernel function used for classification with SVM
Gaussian Process	Used for regression along with classification.
Decision Tree	Performs classification and represents in tree form.

Table 4: *(Continued)*

Machine Learning Models	Explanation
Random Forest	Combination of multiple decision tree.
(Multi Layer) MLP classifier	Artificial neural network supervised learning classifier.
Adaboost	Meta algorithm for classification.
Naive Bayes	Probabilistic bayes based classifier.
(Quadratic Discriminant Analysis) QDA classifier	Technique of Machine learning for classification.

Table 5: Machine Learning Models with their Gained Accuracy

Machine Learning Models	Accuracy Gained
Nearest Neighbor	71.15
Linear Support Vector Machine	77.03
RBF (Radial basis function) Support Vector Machine	65.31
Gaussian Process	67.07
Decision Tree	70.03
Random Forest	74.9
(Multi Layer) MLP classifier	67.6
Adaboost	73.3
Naive Bayes	73.79
(Quadratic Discriminant Analysis) QDA classifier	73.95

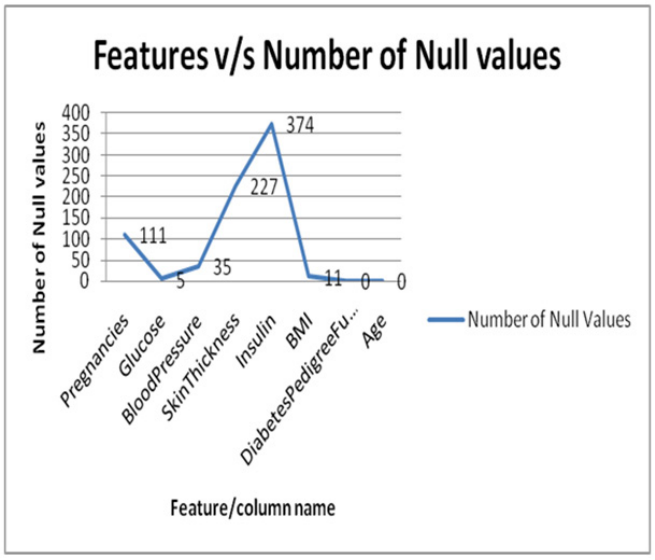

Figure 1: Dataset features v/s Number of Null Values (Before Pre-Processing).

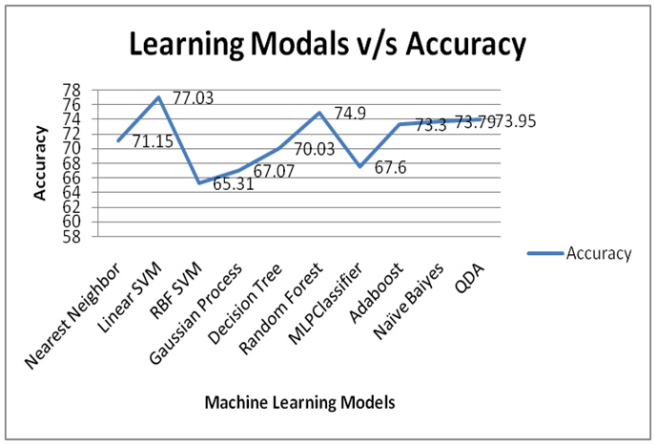

Figure 2: Machine Learning Models v/s Accuracy.